ACEEE Int. J. on Signal & Image Processing, Vol. 4, No. 2, May 2013

Line Spectral Pairs Based Voice Conversion using Radial Basis Function

J.H.Nirmal[1], Suparva Patnaik [2], and Mukesh A.Zaveri[3]
[1,2]Deptartment of Electronics Engineering, [1]K.J.Somaiya College of Engg Mumbai ,India
Email: [1]jhnirmal@engg.somaiya.edu, [2]ssp@eced.svnit.ac.in
[2,3]Deptartment of Computer Engineering, Sardar Vallabhbhai National Institute of Technology, Surat ,Gujarat ,India
Email: [3]mazaveri@gmail.com

***Abstract*— Voice Conversion (VC) is a technique which morphs the speaker dependent acoustical cues of the source speaker to those of the target speaker. Speaker dependent acoustical cues are characterized at different levels such as shape of vocal tract and glottal excitation. In this paper, vocal tract parameters and glottal excitations are characterized using Line Spectral Pairs (LSP) and pitch residual respectively. Strong generalization ability of Radial Basis Function (RBF) is utilized to map the acoustical cues namely, LSP and pitch residual of source speaker to that of target speaker. The subjective and objective measures are used to evaluate the comparative performance of RBF and state of the art GMM based voice conversion system. Objective measures and simulation results indicate that the RBF transformation model performed better than GMM model. Subjective evaluations illustrate that the proposed algorithm maintains target voice individuality, naturalness and quality of the speech signal.**

***Index Terms*— Voice Conversion, Line Spectral Pairs, Long Term Prediction, Gaussian Mixture Model, Radial Basis Function.**

I. Introduction

VC is a process of adapting the acoustical characteristics of a source speaker according to that of target speaker. VC finds it's applications in areas such as personification of text to speech synthesis; audio based learning tool, audition test, audio customization and audio dubbing. [1][2][3]. VC is an exciting new alternative method for building synthetic voices. It can be also used for modifying or normalizing speaker identity as well as modifying a set of acoustic and prosodic characteristics [4]. Voice transformation should be performed from source speech to target speech without losing or modifying the original speech content [5]. VC is performed in two phases: the first one is a training phase, in this phase the speech features of both source and target speaker are extracted and appropriate mapping rules are generated for transforming the parameters of the source speaker onto those of the target speaker. In the second testing phase, the mapping rules developed in the training stage are used to transform the features of source voice signal in order to posses the characteristics of the target voice [6]. The vocal tract and prosodic parameters are modified using signal processing algorithms. There are many applications where intra gender and inter gender voice conversion is required. Many algorithms have been reported in the literature for intra gender and inter gender voice conversion [7].

Atal *et.al*. [8] successfully simulated a female voice using a male voice by rescaling pitch period, formant frequencies and formant bandwidths. In voice conversion, the quality and intelligibility of the synthetic speech depends on time and spectral expansion, compression, pitch modifications, the glottal excitation shape and also on the reproduction of voiced speech [9]. Initially, Abe *et al.* [10] have used vector quantization of the short-time spectra of the speakers, followed by codebooks mapping for voice conversion. Codebook represents the source-target correspondence. The problem with vector quantization and codebook is that the discontinuities are produced due to the hard partition of the acoustic cues which causes degradation of the quality of speech. Fuzzy based vector quantization and mapping techniques have been proposed in [11] which minimises the discontinuities in vector quantization. Different types of codebook mapping methods such as STASC also have been studied in [12][13][14]. The formant frequencies and spectral tilt modification using piece wise linear vector quantization and codebook mapping have been proposed in [15]. Valbret *et al.* [16] has observed that an optimal transformation depends on acoustical characteristics of sound which is to be converted using dynamic frequency warping (DFW). DFW changes only the frequency scale of the source spectrum with a trained mapping function for its corresponding acoustic class. Though, the method proposed in [16] provides high-quality converted speech maintaining its naturalness, the spectral shape of source could not be mapped completely to that of target as only formant frequencies are moved to other frequencies without altering their relative intensity resulting into the loss of intelligent information like loudness of the speech. The further improvements related to frequency-warping have been presented in [17]. Stylianou *et. al*. [18] have used Gaussian mixture models (GMMs) to partition the acoustic space of the source speaker into overlapping classes called as soft partitions. Using these soft partitions, a continuous probabilistic linear transformation function for vectors is obtained. This transformation function contained a parametric representation of the spectral envelope. Kain et.al modified the GMM approach in [19]. However, the quality and naturalness of the converted speech signal is found inadequate due to reconstruction of speech signal using the large number of parameters and it results into over smoothing. Many researchers have provided the solution to GMM based reconstruction of speech signal using different methods like strong vocoding technique, namely, STRAIGHT [20]

DOI: 01.IJSIP.4.2.1168

and post filtering and accurate phase reconstruction based method [21]. Researchers have also provided the solution for over smoothing problem like hybridization of GMM with frequency warping [22] and GMM with codebook mapping [23].

Desai *et.al.* [24] have compared the performance of the ANN and GMM and it is reported that the ANN performs better than GMM. In [25], it is reported that the structural GMM performs better than GMM. GMM with GA based method [26] has been developed using LSP and pitch and shown that it performs better than the conventional GMM based approach. The conversion function has been proposed using the probabilistic model based on inter-speaker dependencies and cross correlation probabilities between the source and target speaker [27].Chen *et.al* have proposed improved method [28] for VC, in which the input frames are divided into voiced and unvoiced frames. The voiced frames are mapped from source to target using GMM and unvoiced frames are stretched or compressed for conversion based on the ratio of vocal tract length (VTL) of source to target. Artificial neural network has been used for VC to exploit the nonlinear relationship between the vocal tract shape of source and target speaker [29][30] [31].

The approach proposed in this paper differs from various methods reported in the literature as below:

a. LSP and fundamental frequency (Fo) are used to extract the source and target features of parallel set of data. Using these features the RBF based neural network is trained for an appropriate mapping function that transforms the vocal spectral and glottal excitation cues of the source speaker in to target speaker's acoustic space.
b. The proposed VC system has been evaluated using both the subjective evaluation and the objective measures like mean square error, pitch and formants.

Our method is compared with conventional GMM based voice conversion approach. The outline of the paper is as follows, section II explains the proposed voice conversion algorithm. GMM based transformation is explained in section III. RBF based transformation model is explained in section IV. Section V and VI describe the results and evaluation using subjective and objective measures respectively. Finally, conclusions are discussed in section VII.

II. Proposed Algorithm For Voice Conversion

The proposed algorithm consists of two phases. In first phase, we extract the LSP and pitch residual based features from source and target speaker data. It is followed by the second phase where we use RBF neural network and GMM which is trained to learn the nonlinear mapping function for source to target speech transformation using the features extracted in the first phase.

A. LSP Based Proposed Voice Conversion

LSPs are an alternative to LPC, as the LPC has many problems such as stability check, quantization and interpolation [34]. LSP are popular due to its excellent quantization characteristics and consequent efficiency in representation.When the LSP are in ascending order in the range [0,1], the resulting filter is guaranteed to be stable. When two LSP values are close to each other, a spectral peak is likely to occur between them which is useful for tracking formants and spectral peaks. The LSP has pure frequency values so it is easy to adapt a perceptual representation such as the Bark scale [35].

The proposed algorithm translates the vocal frequencies of source speaker into target speaker using neural network in two modes. In the training mode the source and target speech is normalized to some predetermined amplitude range and the pitch information is extracted to produce a residual, residual contains vocal tract information which is modeled by LPC. Applying the LPC analysis filter to the residual will result in the vocal tract information being removed leaving lung excitation signal. LPC produces the results in an unstable synthesis filter [36]. So we convert LPC parameters in to LSP. We have mapped the pitch residual and LSP parameters of source speaker on to target speaker using RBF neural network with spread factor of 0.01 [30] and error threshold of 0.00001 also developed mapping using GMM with 64 number of mixtures for vocal tract and 8 number of mixtures for pitch residual respectively. In the transformation phase the LSP and Pitch residual of test speech samples are projected as an input to the trained RBF and GMM model, the transformed LSP and pitch residuals are obtained. To resynthesize the speech signal, LSP parameters are reconverted into LPC, transformed speech is reconstructed using LPC synthesis, as a target speech. The analysis process is essentially reverse of the synthesis process and results in reconstructed speech as shown in figure 1.

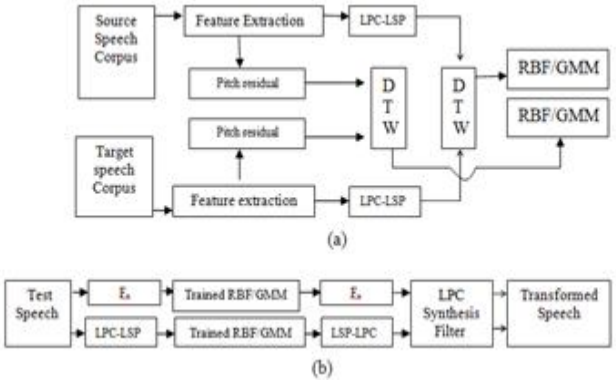

Figure.1: LSP based VC: Training and testing model

III. GMM Based Voice Transformation

The GMM is a soft decision classifier where each class has a Gaussian distribution, The GMM assumes that the probability distribution of the observed parameters takes the following parametric form.

$$p(x) = \sum_{i=1}^{m} \alpha_i N(x, \mu_i \ \Sigma i) \quad (1)$$

Where m is the no of mixture models $N(x, \mu, \Sigma)$ denotes

DOI: 01.IJSIP.4.2.1168

the p dimension distribution with mean (μ_i) and covariance matrix (Σ_i) defined by

$$N(x,\mu,\Sigma) = \frac{1}{\sqrt{2\pi^p \Sigma}} e^{\frac{1}{2}(z-\mu)^T \Sigma^{-1}(z-\mu)} \quad (2)$$

The weighting factor α_i is prior probability of class i with constraints $\sum_{i=1}^{m} \alpha_i = 1 \quad \alpha_i \geq 1$

The input vector x_i assume to be independent. The conditional probability that a given observation vector x belongs to the components Ci of the GMM is given by Bayes rule.

$$p\left(\frac{ci}{x}\right) = \frac{\alpha_i N(x,\mu_i \Sigma_i)}{\sum_{j=1}^{m} \alpha_j N(x_j \mu_j \Sigma_j)} \quad (3)$$

The parameter α, μ, Σ can be estimated with the expectation maximization (EM) algorithm. The EM algorithm is an unsupervised learning in which the component information is unavailable. The EM algorithm is to find the parameters in the GMM that gives maximum likelihood of observed data X. The parameters of the conversion function are estimated by joint density of source and target features. A joint vector $Z=[x^T, y^T]^T$ is used to estimate GMM parameters, where x and y are the aligned source and target speech feature vector.The following parametric form is assumed for the conversion function [32] .

$$F(x)=\sum_{q=1}^{m} p\left(\frac{c_q}{x}\right) [\mu_q^y \Sigma_q^{yx} \Sigma_q^{xx^{-1}}(x-\mu_q)^x] \quad (4)$$

Where $p\left(\frac{c_q}{x}\right)$ is the conditional probability

$$\Sigma_q = \begin{bmatrix} \Sigma_q^{xx} & \Sigma_q^{xy} \\ \Sigma_q^{yx} & \Sigma_q^{yy} \end{bmatrix} \quad (5)$$

$$\mu_q = \begin{bmatrix} \mu_q^x \\ \mu_q^y \end{bmatrix} \quad (6)$$

Where m is no of GMM components. For GMM based mapping, the number of mixtures varies from 2 to 64 dependent on the amount of training data, however it has been observed that, the 64 number of mixtures can characterized the vocal tract of the speaker, whereas the 8 number of mixtures are used to characterized the glottal excitation. The trained transformation model developed during the training is used to predict the vocal tract and glottal excitation of the target speaker. GMM based VC method can effectively convert the characteristics of the speech signal; however the converted speech quality is detorated by excessive smoothing of converted spectra.

IV. RBF BASED TRANSFORMATION

ANN is a huge number of highly interconnected processing neuron. The collective behaviour of ANN, like a human intelligence, demonstrates the ability to learn, evoke and generalize from training patterns. The characteristic of each node is simple. ANN with numerous nodes have powerful self-organizing and self-learning ability with high-quality tolerance and prediction [31]. These characteristics make ANN highly suitable for voice conversion. In proposed method, we have used RBF neural network for transformation. RBFs are feed-forward network consisting of a hidden layer of radial kernels and an output layer of linear neurons as shown in the figure 2. The two layers in an RBF carry completely different roles. The hidden layer performs a non-linear transformation. The output layer performs linear regression to predict the desired targets. Each hidden neuron in an RBF is tuned to local region of feature space using a radially symmetric function. Hidden unit activation is determined by distance between the input vector X and a prototype vector μ.

$$\emptyset_j(X) = f(|X-\mu_j|) \quad (7)$$

The most commonly used activation function is Gaussian kernel defined as,

$$\emptyset_j(X) = exp\left[-\frac{[|X-\mu_j|]^2}{2\sigma_j^2}\right] \quad (8)$$

The activation of the output unit is determined by dot product between the hidden activation vector Õ and weight vector w as,

$$y_k = \sum_{j=1}^{N_H} w_{ij} \emptyset_j (|X-\mu_j|) \quad (9)$$

The output unit performs weighted sum of hidden units. The RBF neural network is trained to map LSP as a vocal tract parameters and pitch residual (Fo) as glottal parameters between source and target speaker. A generalized learning law is used to adjust the weights of this neural network so as to minimize the mean squared error (MSE) between the desired and the actual output values.

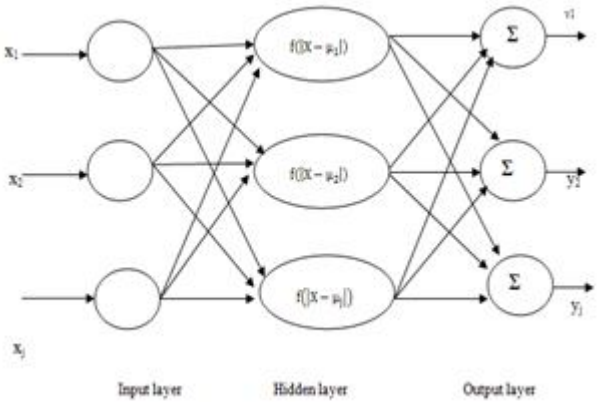

Figure 2: Radial Basis Function Neural Network

For training, the LSP of the source speaker voice are used as input and LSP of target speaker voice are used as a output of the network. The weights of the hidden and output layers of the networks are adjusted in such a way that the source speaker's patterns are converted into the target speaker's patterns. The training step is a supervised learning procedure.

DOI: 01.IJSIP.4.2.1168

V. Simulation And Results

The proposed algorithm is evaluated using large number of databases. These databases consist of standard database as well as our own database consisting of Gujarati and Marathi regional Indian languages. Our own database consists of 48 sentences recorded using a high quality microphone (Sony V_120). The speech samples are recorded at a frequency range of 16 KHz; the speakers are trained well before capturing the speech corpus. The recorded speech samples are processed labelled and stored. Our database is parallel database; that is source and target speakers are requested to narrate the same sentence. We call it as parallel database. These sentences are collected from five males and five female's speakers.

We have performed the VC for intra gender and inter gender using proposed algorithm. For this, we have evaluated the performance of our algorithm using standard databases, namely, CMU ARCTIC database consisting of utterances recorded by 7 speakers. We successfully transformed SLT (U.S. female) to BDL (U.S. male) and BDL (U.S. male) to SLT (U.S. female) which is described later in this section. Our algorithm is able to perform the mapping of the LSP, pitch residual, of source to that of target speaker using RBF and GMM based mapping with high accuracy.

Figures 3 represent the simulation results using GMM based method. The sample results of our RBF based speech conversion algorithm for inter gender speaker are shown in the figures 4. In figures 3-4, the left column represents the speech signal waveform of source, target and transformed in order. The right column in the figures 3-4 displays the spectrogram of source, target speaker and transformed speech in order from top to bottom. The spectrograph is a two dimensional time and frequency graphical plot of the energy present in the signal. Figure3, display the waveforms and spectrograms for female to male voice conversion where as Figure 4 displays the results for male to female voice conversion. Similarly figure 3 and figure 4 depict the result for female to male and male to female voice conversion for specific sentence uttered from Gujarati Indian regional language. From the visual perception of the figures it is clearly seen that the spectrogram of the target speaker is similar to that of the transformed speech. The performance of the ANN based transformation is better than GMM based transformation. The proposed algorithm is able to transform source speech to target speech successfully.

VI. Objective And Subjective Evaluations

In this section we evaluate our algorithm based on objective and subjective parameters. We use two objective based evaluation parameters (i) mean square error (MSE) and (ii) similarity measure using pitch and formant. We perform similarity based evaluation using pitch and formant because pitch represents the fundamental frequency of the speech and formant represents resonant frequency of vocal tract. This similarity measure allows us to differentiate between male to female and female to male voice conversion.

A. MSE based Objective Evaluation

In this section we provide the objective evaluation for RBF and GMM based VC systems to measure the differences between the target and transformed speech signals. Since many perceived sound differences can be interpreted in terms of differences of spectral features and mean squared error (MSE) are considered to be a reasonable metrics; for both mathematically and subjectively analysis. The MSE between target audio vector *p* and transformed audio vector *s* is calculated as per below equation on a sample by sample basis.

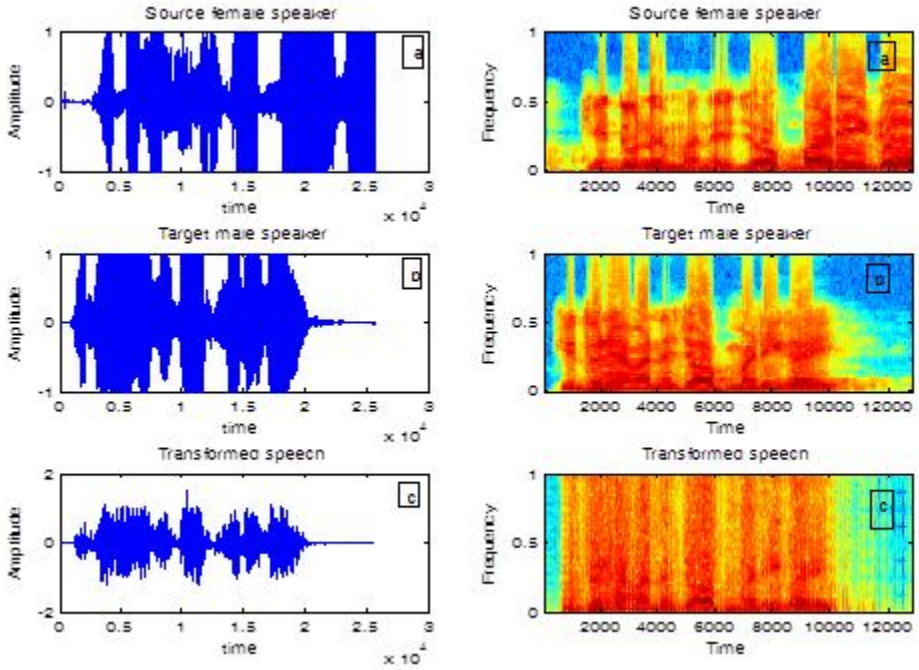

Fig 3. a) Source, b) Target and c) Transformed speech of same sentence waveform for a female-to-male speaker using GMM based speech transformation

DOI: 01.IJSIP.4.2.1168

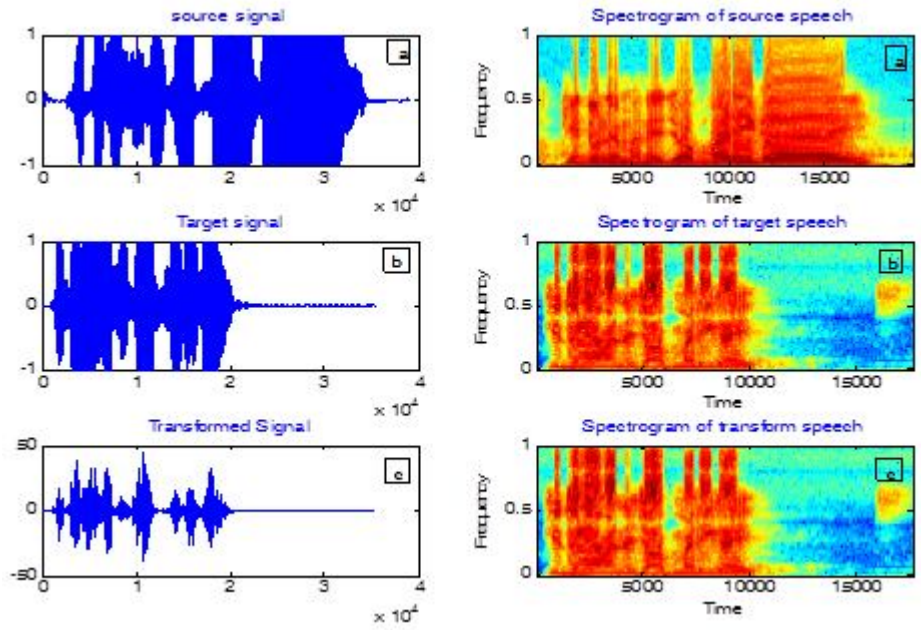

Fig 4. a) Source, b) Target and c) Transformed speech of same sentence waveform of a male-to-female speaker of RBF based speech transformation

The average square difference between two vectors is used to evaluate the objective performance of mapping algorithms shown in Table I.

$$E = \frac{1}{N}\sum_{i=0}^{N-1}[s(i) - p(i)]^2 \quad (10)$$

TABLE I: MSE FOR GMM AND RBF BASED VC SYSTEMS

	GMM	RBF based
Male to Female	0.0206	2.8642e-029
Female to Female	0.0328	7.3574e-030
Female to Male	0.0185	3.6461e-029
Male to Male	0.0763	6.9978e-029

B. Pitch and formant based similarity measures

This similarity measure is used to find out how close the pitch and formants of speech produced by the application with the source and target speaker. In order to evaluate the performance of the RBF and GMM based transformation we have done comparison based on the fundamental frequency (f0) and spectral formant frequencies, the fundamental frequency (f0) of the vocal fold vibration determines the perceived pitch of a voice, most often it is higher in females than in males. The spectral formants determine the characteristic timbre of each speaker's voice, so that the listener can recognize familiar voices and discriminate the gender from familiar or unfamiliar voices [37].A formant is a distinguishing or meaningful frequency component of human speech. R.vergin et.al classified male/ female based on the location of the first and second formants in the frequency domain [38]. This classification is used to split up automatically the training corpora into male and female speakers. We have calculated the pitch and the formant frequencies of source, target and transformed speech and shown in following tables.

TABLE II: PITCH AND FORMANTS OF SOURCE, TARGET MALE-FEMALE TRANSFORMED SPEECH

	Source	Target	Transformed Signal	
	Male	Female	GMM	RBFN
Avg pitch	192.71	317.56	231.12	325.658
Min Pitch	104.42	240.6	86.342	238.581
Max Pitch	268.8	496.5	395.07	499.0211
Ist Formant	545.6	659.7	520.6	961.888
2ndFormant	1826.1	1924.8	1768.7	2058.4
3rdFormant	2797.62	2926	2429.1	3093.90

As shown in table 2 the average pitch of the target speaker (female) is larger than the source speaker (male). RBF based Transformed speech signal's average pitch is more than pitch of source signal and closed to the average pitch of the target speech as compared to pitch of GMM based transformed speech.

As per the table 3 the average pitch value of the target speaker (male) is less than the source speaker (female). The RBF based transformed speech signal's average pitch value is less than pitch of source signal and closed to the average pitch value of the target speech. The fundamental frequencies (pitch period) of the women are higher than the men, as can be observed from table 2 and 3 the target speech is similar like the transformed speech and it is clearly transformed from men to women and vice versa. It is also shown that the line

DOI: 01.IJSIP.4.2.1168

TABLE III : PITCH AND FORMANTS OF SOURCE, TARGET AND FEMALE TO MALE TRANSFORMED SPEECH

	Source Sig	Target sig	Transformed Signal	
	Female	Male	GMM	RBF
Avg pitch	372.5	192.71	75.69	189.6
Min Pitch	234.5	104.425	74.42	102.8
Max Pitch	425.56	268.851	77.56	268.1
Ist Formant	698.16	512.450	577.144	1021.8
2nd Formant	2037.5	1746.36	1488.6	2002.6
3rd Formant	3008.0	2787.4	2273.1	3217.7

spectral pair based technique performed better than GMM based transformation.

C. Subjective evaluation

Evaluation of the synthesized speech of desired speaker using RBF and GMM based VC systems by standard subjective test can be attributed due to inefficiency of objective evaluation techniques in defining good objective distance measures which are perceptually meaningful. In this paper two subjective evaluation tests have been discussed namely i) Mean Opinion Score ii) ABX test

1.Mean Opinion Score

To evaluate the overall accuracy of the conversion, a listening test is carried out for evaluating the similarity between the converted voice and the target voice to find the performance of GMM based transformation against RBF based transformation. 9 listeners give scores between 1 and 5 for measuring the similarity between the output of the two VC systems and the target speaker's natural utterances. The results of this MOS similarity tests are provided in figure 5, which indicates that the RBF and GMM based voice conversion systems completely characterized the characteristics of the source speaker. The RBF based voice conversion has slightly more resemblance to target speech as compared to the GMM based VC system.

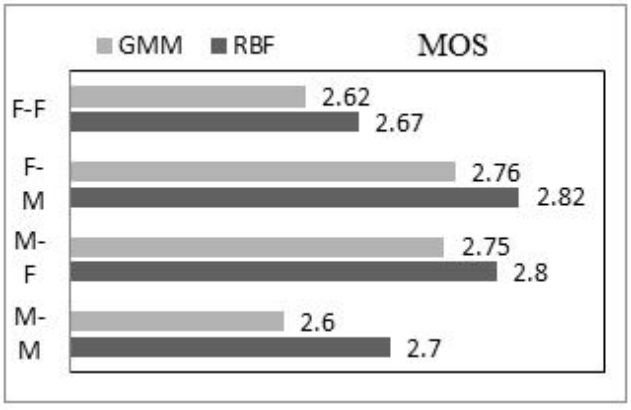

Fig 5: Mean Opinion Score for RBF and GMM based VC Techniques

2.ABX test

The ABX test is an alternative perceptual test to evaluate the performance of VC systems, which is often used to find the similarity between, converted and target voice. In this test, independent listeners have to judge whether a given utterance X is similar to utterance of speaker A or B, where X is a transformed speech and A and B are source speaker and target speaker speech respectively. Utterances of source and target speakers are taken from corpus. It is presented to independent listener in random order. In total 20 utterances are tested and shown in figure 6. This 9 college student listeners are asked to rate the transformed speech on a scale of 1 to 10, the lower value representing a non recognizable converted speech and vice versa.

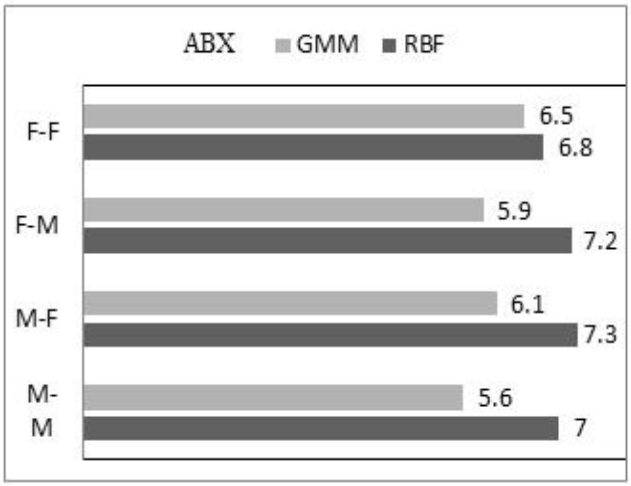

Fig 6 : ABX test for RBF and GMM based VC Techniques

Subjective evaluation is an effective technique to assess the quality and naturalness between the transformed and target speech. The basic reasons for variation in MOS and ABX values are variation in gender, shape, size of vocal tract and glottal. It also varies with respect to the position of teeth, lips, tongue, jaw and velum.

CONCLUSION

In this paper, we have proposed a novel technique using LSP as spectral features and pitch residual as glottal features. The RBF based and GMM based mapping functions are developed to map the acoustical cues of the source speaker according to that of the target speaker. The inter gender and intra gender VC is performed using proposed algorithm .We have experimentally verified that, the fundamental frequency range of female speech is higher than male speech and vice versa. The comparative performance of RBF and GMM based models are studied using objective and subjective measures. The evaluation results indicate that RBF can be used as an alternative to the GMM based transformation model. The subjective evaluation convinced that the quality and naturalness of the transformed speech can be achieved with the proposed algorithm.

REFERENCES

[1] A.Kain,:'High resolution voice transformation', PhD dissertation, Oregon Health and Science University, 2001.

DOI: 01.IJSIP.4.2.1168

[2] Daniel erro eslava,:'Intra-lingual and cross-lingual voice conversion using Harmonic plus Stochastic models',PhD dissertation, universitat politècnica de catalunya,2008.

[3] O.Turk,:'Cross-lingual voice conversion', PhD dissertation, Bogazii University,2007.

[4] Stylianou,Y.,'Voice Transformation: A survey'. International Conference on Acoustics, Speech and Signal Processing, ICASSP 2009 pp 291-298.

[5] Anderson Fraiha, Machado: 'Techniques for Cross lingual Voice Conversion'. IEEE International Symposium on Multimedia 2010 pp 365-369.

[6] K.Sreenivasa Rao.,:'Voice Conversion by Mapping the Speaker-specific features using Pitch Synchronous Approach', Computer Speech and Language, Elsevier, July 2010 Vol. 24, pp. 474- 494.

[7] M. R. Schroeder,:'Vocoders: Analysis and Synthesis of Speech.' Proc. IEEE (54),1966 pp. 720-73l.

[8] B. Atal, and S. Hanauer,.' Speech Analysis and Synthesis by Linear Prediction of the Speech Wave', JASA.1971 (50), pp. 637- 655.

[9] Childers, D.,Yegnanarayana,B., Ke Wu ,:'voice conversion: factors responsible for quality', Proc.Int. Conf. on Acoustics,Speech,and Signal. March 85.pp 748-751

[10] Abe, M.,Nakamura, S., Shikano, K. Kuwabara, H.:'Voice conversion through vector quantization' International Conference. Acoustics, Speech, and Signal Processing,ICASSP-88.pp 655.

[11] Abe M A :'Segment-Based Approach to Voice Conversion', International .Conference Acoustics, Speech, And Signal processing, ICASSP-91.,PP 765.

[12] Arslan, L.M.,Talkin, D.,:'Voice conversion by codebook mapping of line spectral Frequencies and excitation spectrum'. International. Proceedings EUROSPEECH, Rhodes, Greece, Vol. 3, pp.1347-1350.

[13] Shikano, K,Nakamura S,Abe M,:'Speaker adaptation and voice conversion by codebook Mapping' IEEE International Symposium on Circuits and Systems, 1991,,vol 1,pp.594-597.

[14] L. M. Arslan,: 'Speaker transformation algorithm using segmental codebooks', (STASC) Speech Communication, 1999, 28(3): 211-226.

[15] Hideyuki Mizuno, Masanobu Ab.:'Voice Conversion Based On Piecewise Linear Conversion Rules Of Formant Frequency And Spectrum Tilt'. IEEE International conference Acoustics, Speech, and Signal Processing 1994 pp 469-471.

[16] H. Valbret, E. Moulines, J.P. Tubach ,:'Voice Transformation Using PSOLA Technique'.Acoustics, Speech,and Signal Processing,ICASSP-92 pp I 145-148.

[17] Z.W. Shuang, R. Bakis, S. Shechtman, D. Chazan, and Y. Qin.:'Frequency warping based on mapping Formant parameters'. in Proc. International .Conference. Spoken Lang. Process., 2006.pp

[18] Stylianou, Olivier Cappa,:',A system for Voice Conversion based on probabilistic Classification and Harmonic plus Noise Model 'International Conf. Acoustics, Speech and Signal Processing, Proceedings of the 1998 .pp 281-285.

[19] A.Kain, and M.W. Macon,:'Spectral voice conversion for text-to-speech synthesis',Proc International conference Acoustic Speech and Signal processing ICASSP,Seattle U.S.A., pp. 285-288, May 1998.

[20] Toda, T.; Saruwatari, H.; Shikano, K.,:'Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum' International Conference on Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01).pp 841-844.

[21] H.Ye and S. Young, 'Quality-enhanced voice morphing using maximum likelihood transformations', IEEE Trans. Audio, Speech, Lang. Process., vol. 14, no. 4, pp. 1301–1312, Jul. 2006.

[22] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano,:'Maximum likelihood voice conversion based on GMM with straight mixed excitation'. in Proceeding . Interspeech, 2006.pp 2266-2269

[23] Daniel Erro,Asunción Moreno, And Antonio Bonafonte,:' Voice Conversion Based On Weighted Frequency Warping,",IEEE Transactions On Audio, Speech, And Languag Processing,Vol.18,(5),July 2010 pp 922-931.

[24] S.Desai,E.V.Raghavendra,B.Yegnanarayana, Alan Black,and K. Prahallad.,:'Voice conversion using artificial neural networks', in Proceedings of IEEE Int.Conf. Acoust., Speech, and Signal Processing, 2009 .pp 3893-3897.

[25] Xie Chen, Wei-Qiang Zhang, Jia Liu Xiuguo Bao,:An Improved Method for Voice Conversion Based on Gaussian Mixture Model'. International Conference on Computer Application and System Modelling (ICCASM 2010) PP V4-404-408.

[26] Chen Zhi ,Zhang Ling-hua.,:'Voice Conversion Based on Genetic Algorithms'.International .Conference on. Communication technology 2010 pp 1407- 1410.

[27] Y.Stylianou,O.Cappe,and E. Moulines,:'Statistical methods for voice quality transformation," in Eurospeech, 1995, pp. 447–450.

[28] Xie Chen, Wei-Qiang Zhang, Jia Liu Xiuguo Bao:' An Improved Method for Voice Conversion Based on Gaussian Mixture Model'. Int. Conf.on Computer Application and System Modelling (ICCASM 2010) PP V4-404-408.

[29] M.Narendranath, H. A. Murthy, S. Rajendran, and B. Yegnanarayana, 'Transformation of formants for voice conversion using artificial neural networks', Speech communication, vol. 16, pp. 207–216, 1995.

[30] J.H.Nirmal, S.S.Patnaik,M.A. Zaveri, "Voice Transformation using Radial Basis Function",Third International Conference on Recent Trends in Information, Telecommunication and Computing ITC 2012, Springer-Verlag Berlin Heidelberg (2012) pp 271-276.

[31] C. Orphanidou, I.M. Moroz, S.J. Roberts,'Wavelet-based voice morphing'. WSEAS J. Syst. 1 (3) 2004 pp 3297–3302

[32] R.H. Laskara, D. Chakrabartyb, F.A. Talukdara, K. Sreenivasa Raoc, K. Banerjeea, "Comparing ANN and GMM in a voice conversion framework", Applied Soft Computing 12 (2012) 3332–3342.

[33] Rodrigo Capobianco Guidoa,Lucimar Sasso Vieiraa, Sylvio Barbon Juniora,,'A Neural Wavelet Architectures for voice Conversion',,ScienceDirect Neurocomputing,71 (2007) pp 174-180.

[34] S. Grassi, A. Dufaux, M. Ansorge, and F. Pellandini,' Efficient algorithm to compute LSP parameters From 10th-order lpc coefficients'. Int.Con. on Acoustics, Speech, and Signal Processing (ICASSP97) 3, 1707-1710, 1997

[35] Lan Vince McLoughlin.,',Line Spectral pairs'.,Elesevier Signal Processing (2008) pp 448-467

[36] Lan Mcloughlin,'Applied speech and audio processing with matlab examples' Cambridge Publication.,first edi. 2009.

[37] Rivarol Vergin,Azarshid F,Doughlough shahguansy., Robust Gender Dependent Acoustic Phonetic Modeling in continous speech recognition based on new automatic Male Female classification'.Int.Conf. Spoken Language Processing

DOI: 01.IJSIP.4.2.1168

ICSLP 2006 pp 1-4

[38] Pawan Kumar, Nitika Jakhanwal, Anirban Bhowmick, Mahesh Chandra.,'Gender classification using pitch and formant'. International.Conference. on Communication, Computing & Security 2011 pp 319-324.

Jagannath Nirmal received his B.E. and M.Tech. degrees in Electronics Engineering from SGGSIE&T, Nanded and VJTI, Mumbai in 1998 and 2008 respectively. Currently he is pursuing Ph.D. in Speech Processing at SVNIT, Surat. His main research interest includes Speech Process ing, Patterns Recognition and Classification, Adaptive filtering ans Signal Processing.

Suprava Patnaik received the Ph.D. degree in electronics engineering from Indian Institute of Technology, kharagpur, India in 2003. Presently she is guiding Ph.D. students in the area of Signal and Image processing. She is the author of many articles in reputed journals and conferences. Her main research interests are wavelet transform, pattern recognition and classification, machine learning, artificial intelligence, signal and image processing.

Mukesh Zaveri received the Ph.D. degree in electrical engineering from Indian Institute of Technology, Bombay, India in 2005. Presently he is guiding Ph.D. students in the area of Wireless communication, Signal and Image processing. He is the author of many articles in reputed journals and conferences. His main research interests are wireless networking, computer vision, machine learning, artificial intelligence, signal and image processing.

DOI: 01.IJSIP.4.2.1168